"A Multi-Method Exploration of Crime Hotspots"

Software Evaluation: SpaceStat

By
Joseph Szakas

With
Nancy La Vigne, Eric Jefferis, Cyndy Nahabedian
Elizabeth Groff, Julie Wartell and Maureen O'Connell
Crime Mapping Research Center, National Institute of Justice[1]

Presented at the
Annual Meeting of the *Academy of Criminal Justice Sciences* (ACJS)
Albuquerque, New Mexico
March 11, 1998

[1] Points of view are those of the author and do not necessarily represent the view of the U.S. Department of Justice or the National Institute of Justice.

"A Multi-Method Exploration of Crime Hotspots"
Software Evaluation: SpaceStat

Introduction:

The analyses of spatial data and subsequent conclusions derived from these analyses have always been sources of concern for the decision-maker. These concerns range from the accuracy of the spatial data (e.g. data acquisition, timeliness, positional accuracy, and attribute accuracy), to the appropriateness of the selected algorithm for the specific task at hand (i.e., should this spatial operation be performed?), to the accuracy of the algorithm chosen (i.e., is the algorithm properly implemented?). It therefore comes as no surprise that the many methods and software packages that exist today have some degree of variance in the processing of spatial data. It is the goal of this study to examine one of these methods for the purpose of a comprehensive comparison with several other methods and software packages.

To aid in the comparison, the attempt has been made to minimize, as best as possible, the following parameters: the data, and the study areas. Although data are often modified to meet specifications of the software package, it is hoped that these (duly noted) modifications will not detract from the results. The primary focus of the study should be the algorithms, and the ability of the software package to correctly apply those algorithms. The following discussion is an attempt to describe the use of SpaceStat for hot spot detection. This description comments on user friendliness, validity, and uncertainty of results.

Package Description:

The package under evaluation in this portion of the study is SpaceStat. SpaceStat is a program created by Luc Anselin for the purpose of applying certain advanced statistical operations to spatial data. Unlike many of the other software packages being examined in this multi-method study, SpaceStat is a non-graphical software package. Managing spatial data normally requires the ability, after processing, to visually review the spatial distribution of results for final decision making. SpaceStat provides results in a spreadsheet or tabular format. In order to "see" the results, not in the numerical results but in the numerical result's spatial distribution, additional software is needed. SpaceStat is, as the SpaceStat manual puts it, "loosely connected" to ESRI's ArcView and to other GIS/mapping packages. This "loose" connection allows SpaceStat to focus its processing efforts on the statistical algorithms, and leave the graphics to more established mapping applications (e.g. ArcView, Idrisi, and MapInfo). The price the user pays for this division of labor (not to mention the need to purchase two software packages) is that the user must move the data from one package to the other, which may not be terribly difficult, but can be cumbersome. SpaceStat is a DOS-based application. That means that the user will, depending on the size of the crime analysts' datasets, be either in DOS or in the Windows 95 operating system environment. [Note: for small datasets, a DOS shell can be opened within Windows 95 to provide a more harmonious working environment.]

Method Description:

Within SpaceStat, the statistical operation selected to obtain hot spots was the G_i Statistic. The G_i Statistic is part of a larger family of G statistics developed by Getis and Ord (1992). These statistics are distance-based such that for each point, the statistic is computed using only neighbors within a pre-defined distance (d). These neighborhoods are defined for a specific distance by a weight matrix (W (d)). The formula for the G_i statistic is shown below:

$$G_i = \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j}$$

For each observation, *i*, the G_i statistic sums all the values (x_j) of the neighbors (w_{ij} (d)) of *i* within a distance of d. The results are presented with standardized Z values. The interpretation of the G_i statistic, according to the SpaceStat manual, is as follows:

> "A positive and significant z-value for a G_i, statistic indicates spatial clustering of high values, whereas a negative and significant z-value indicates a spatial clustering of low values." (Anselin, 1995, page 23-2)

Hot spots, although an ill-defined term, can be thought of as a clustering of high values. With this in hand, the G_i statistic appears to be an appropriate selection. It is important to note that SpaceStat does not have a "hot spot" function per se. This appears to be the best statistic within the SpaceStat arsenal to accomplish the stated task.

Procedure:

The spatial data used in this study are a collection of burglaries and robberies that have occurred in Baltimore County from November 1996 through November 1997. The first step involved pre-processing to create four datasets used in this study: robberies for study area 1, robberies for study area 2, burglaries for study area 1, and burglaries for study area 2. Pre-processing was done in a two-stage process. The first pre-processing step organized the data into its four study units. The second process involves manipulating the data to: reformat the data according to SpaceStat specifications; reduce the number of elements in the dataset for SpaceStat to process; and to adjust the data values to meet the requirements for the application of the specific operation (in this case the G_i statistic).

These data are event data with geographic coordinates provided (i.e., latitude and longitude). Within this file there were over 6,000 burglaries and approximately 1,200 robberies with a corresponding geographic coordinate. The data were provided in a DBase III format (de facto standard for database files in the mapping world). Using ArcView, these data were converted into an ASCII delimited text file for processing by SpaceStat. There was no projection of these datasets (i.e., x = longitude and y = latitude). The positional accuracy of this data is unknown. Some manipulation of the actual datasets was needed in order to execute hot spot detection using SpaceStat. The datasets used for this study were broken into two categories: robberies and burglaries. Then subsets were extracted from each category for initial analysis upon a smaller geographical space. The larger area of study, called study area 1, encompassed all of Baltimore

County. The smaller area, called study area 2, focused upon the southeast corner of the county (due east of the city of Baltimore, which is not contained within Baltimore County).

For SpaceStat to process point data using the G_i statistic, there must be a field to apply the statistic. Therefore the data format was modified to include a frequency field for the number of robberies and burglaries respectively. It proved unsuccessful to simply append a new field with constant values (i.e., a field in which each observation had a value of 1). An alternative method was selected to aggregate the data and create a frequency field (i.e., number of robberies, and number of burglaries) for each dataset except burglaries for study area 1. Although this can be considered pre-clustering, this exact method was used to derive the pin map data used in this study. In this respect, the data are the same as were provided to the human crime analyst. For the dataset containing burglaries for study area 1, the number of observations proved too large even after aggregation similar to the pin map data and it was necessary to further reduce the number of elements in the dataset. It should be noted that SpaceStat can indeed process large datasets, but the creation of one portion of the procedure required over four hours of computation time and file sizes approaching 100 megabyte files. The decision was then made that the original dataset, pre-aggregated, be moved to a polygon dataset of census block groups for the county. This point to polygon transfer provided a frequency field for each polygon for the number of burglaries. The geographic coordinates of each point were lost, and all attributes were now tied to the centroid coordinate of each polygon in

the new dataset. Comparison with results from the other studies will be needed to fully comprehend the effects of this decision to aggregate to the block groups.

In the final stage of data pre-processing adjustment of data values was required. Application of the G_i statistic requires that all values within the field being processed be positive (i.e. greater than zero). After aggregation into the polygon set, there may indeed be polygons without any burglaries located within their geographical space. To overcome this, a small constant value (0.1) was added to each value in the number of burglaries field.

Once the datasets were created, the actual data processing began. This process was a cyclical one, in which data begin in ArcView (or some mapping package) are moved into SpaceStat for processing, and then returned to ArcView for visualization and evaluation. Data are moved from ArcView to SpaceStat by exporting the files into a "common" format. This common format is an ASCII delimited text file. One must select which fields to export. Important fields to export are a unique key field, the X and Y coordinate fields, and the field for which the G_i statistic will be processing (in this case the fields containing the number of burglaries and number of robberies respectively). Once this ASCII field is created, it must then be modified (removing the quotation marks and commas from the field and entering the number of observations and the number of fields). This is a relatively straightforward task, and although time consuming and tedious, it does provide the opportunity for a visual examination to insure that all the data

were exported correctly. When there were blank entries in any of the fields to be exported, the subsequent input into SpaceStat would fail.

Data were imported into an internal format and the distance matrix was then created. The distance matrix is actually arc distances, since the data coordinates are in latitude and longitude. The ability to calculate arc distances verses Euclidean distance (normally the only option) is an advantage of using SpaceStat. From the distance matrix, neighborhood bands (called contiguity matrices) are created. The creation of contiguity matrices is a requirement for the G_i statistic. Using a distance band provides that only values within a certain tolerance will be used during the processing of the statistic. The selected statistic (G_i) is applied and the results are written out to a file that can be brought directly into ArcView.

ArcView can then import the results, and a table join operation is performed so that the results of the G_i Statistic can be associated to the points on the map through a unique key field. The results of the G_i Statistic are in z-values. Since only the largest and most significant z-values are of importance to us (as per the SpaceStat manual) only z-values with values larger than or equal to two (i.e., at least two standard deviations away from the mean) will be placed upon the final maps. The analyst or operator can now visually evaluate the results and perhaps repeat the cycle to find a “correct” distance band.

Determination of the distance band is a significant parameter for hot spot detection. The selection of a distance band and the subsequent evaluation of the effectiveness of the distance band are probably the most subjective parameters in the hot spot detection process. The value is determined by size of the unit of analysis (e.g., actual point data, aggregated data, or census tract data). Experience (knowledge of the area, and geographical features) also plays a key role in the success of any hot spot detection process. A detailed flow chart that graphically details the detection of hot spots using SpaceStat appears in Appendix A of this document.

Summary of Findings:

The following are answers to questions posed by the moderator of this investigation.

1. What is the underlying algorithm?

The G_i statistic developed by Getis and Ord. This formula is:

$$G_i = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}$$

This is a statistic that creates a value for each observation i. The statistic is the sum of the values of neighbors of observation *i* divided by the sum of all the other observations. Neighbors are defined by a distance band $w_{ij}(d)$. [Note: the observation being calculated is not part of the statistic.] The user bases this statistic upon a distance band, where the value of the band is defined by the user a-priori. For example, the lower bound equals zero miles, and the upper bound is equal to d miles. The distances used by

SpaceStat to create the distance bands are arc distances since the data were in latitude and longitude. The formula for the arc distance also appears in the SpaceStat menu, assuming a fixed radius of the earth. Many packages can only provide Euclidean distances that are not as accurate for determining distances on the earth's surface, so this is a distinct advantage of using SpaceStat. The values of the x_j's must be positive values, Although this positive requirement doesn't appear to be too rigid, when the data are aggregated to census blocks, some census blocks may indeed have zero crimes. A small constant value was added throughout the entire data set to correct for this shortcoming. The results will have a slight shift in the mean, but that shift will be compensated for in the z-values.

How user friendly is the program?

The program requires detailed knowledge in both ArcView and SpaceStat. The learning curve for SpaceStat alone is steep. True of most powerful statistical software packages, SpaceStat is more concerned with providing tools for spatial data analysis, than the appeasement of novice users. As mentioned before, SpaceStat is a DOS based program. With a DOS system comes the filename limitation of eight characters, which is a major limitation within SpaceStat. As new files are produced, any filename over eight characters will be truncated. With ArcView running under Windows 95, there is some loss of productivity in the moving of data from one package into another. With ArcView exported files, editing was still required for input into SpaceStat. Output from SpaceStat did not require any editing, but a unique key was needed for the table join to match calculated statistics with their spatial object in the map. SpaceStat has addressed some of

these concerns through the use of ArcView extensions, which were not used during this test, but the dual operating system platform issue leaves something to be desired.

Do the resulting hot spots have face validity?

The hotspots do have face validity, especially for both robberies and burglaries in study area 2, and for robberies in study area 1. Burglary hot spots in study area 1 are suspect because of the use of block group centroids. Large block groups may skew the distance calculations, hindering the effectiveness of the distance function. Also, boundary conditions around the city of Baltimore and around Baltimore County itself is an issue not dealt with by SpaceStat. Most of the hot spots are somewhat clustered about the city, which is expected.

Does the program have practical utility?

SpaceStat is a package that represents an advanced toolbox more appropriate for the statistical researcher than for the police crime analyst. Indeed, after the examination of all the pre-requirements in order to do hot spot detection in SpaceStat, one has to wonder if anyone else is doing hotspot detection "correctly." The fact that implementation of SpaceStat and ArcView requires two different operating systems almost excludes it from having practical utility.

Is the program flexible?

Without belaboring the point, SpaceStat is DOS based and ArcView is Windows 95 based. The words "loosely connected" basically mean that there is work for the user

to move data. SpaceStat focuses solely on spatial statistics and leaves the graphics to ArcView. This keeps with the old adage of not re-inventing the wheel. Within SpaceStat there are a host of tools, and this investigation only scratched the surface. Statistics such as Moran's I, Geary's C, scatter plots, box plots, variograms, other G statistics, and higher ordered contiguity matrices give SpaceStat a very powerful and sophisticated collection of spatial statistics, and it is nice to be able to find all of these statistics within a single package.

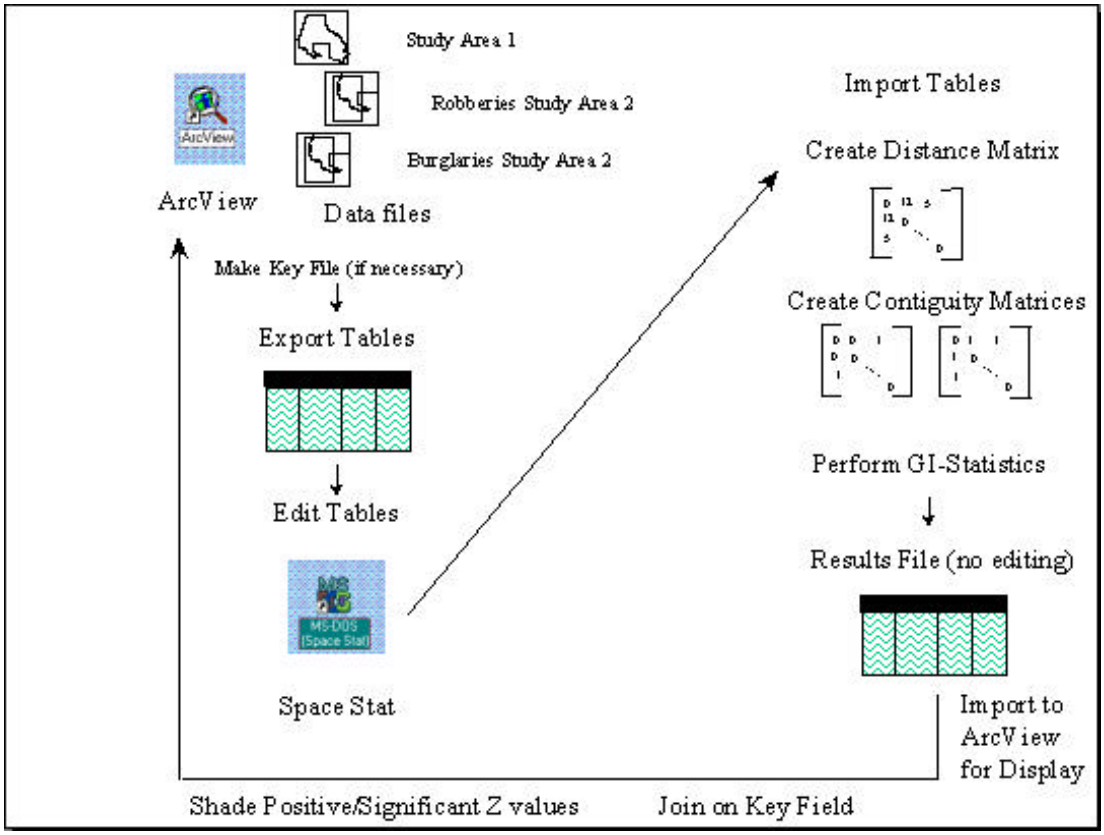
What did you like best and least about this product?

SpaceStat, once understood, is a very powerful package. It definitely allows the user a large portion of control and is up front with all formulas used. The cyclical nature of moving data back and forth from SpaceStat to ArcView was cumbersome. The fact that SpaceStat is able to do this in-depth level of analysis is very impressive.

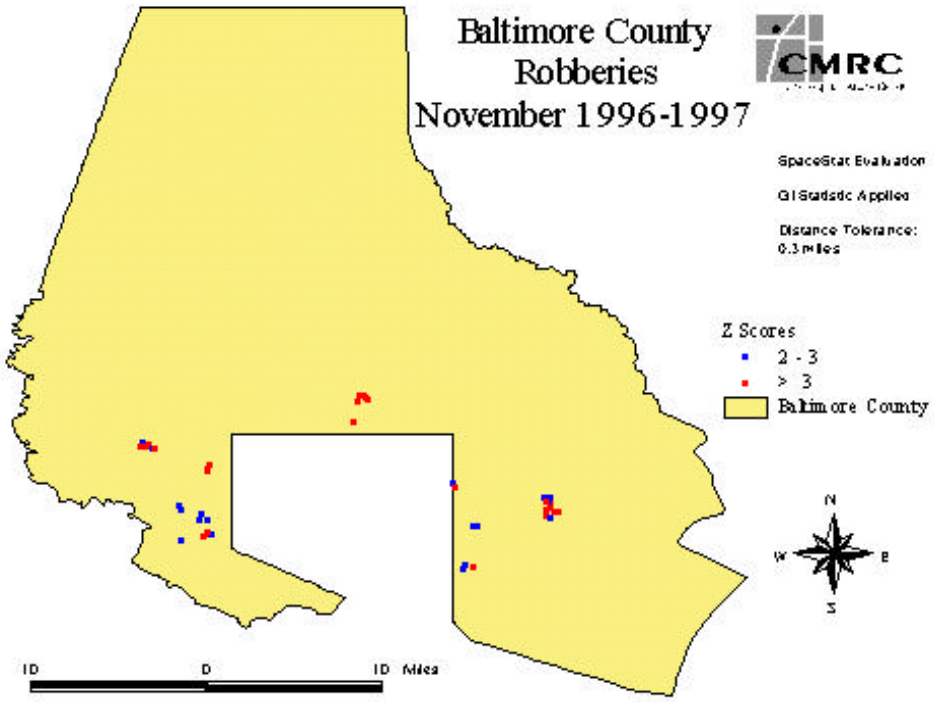
Conclusion:

For the serious spatial statistician, SpaceStat is a dream come true. For the novice user with only a manual in hand and a limited knowledge in statistics, that dream turns into a nightmare. Once the statistic is chosen, the actual implementation of the software is simply a few keystrokes. SpaceStat certainly has a place in the research community. It demonstrates the ability to identify hot spots, and goes towards the next step, which is to understand the factors behind hot spots.

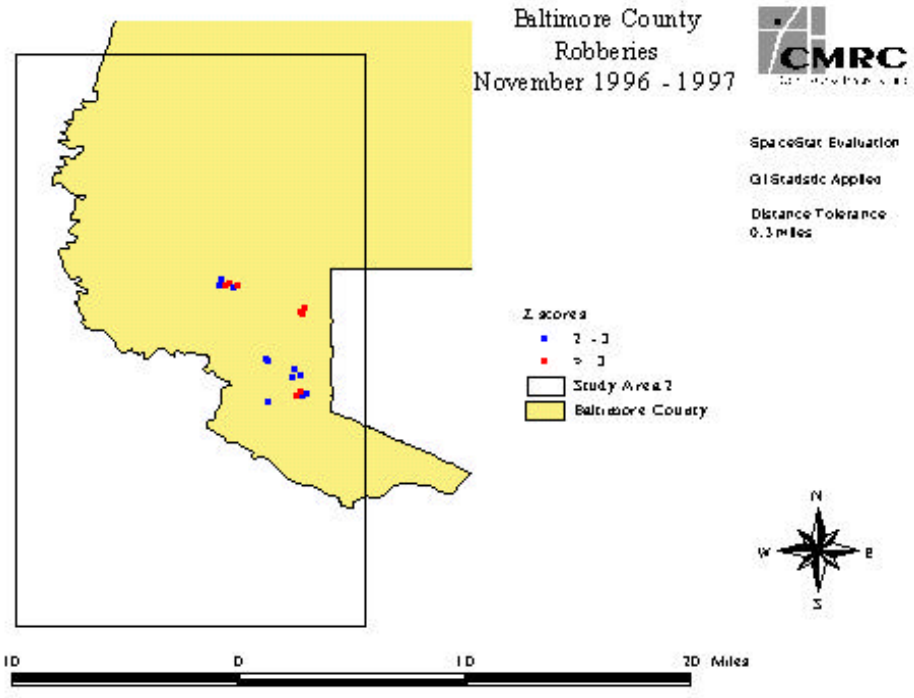
Appendix A: Data Flow Model



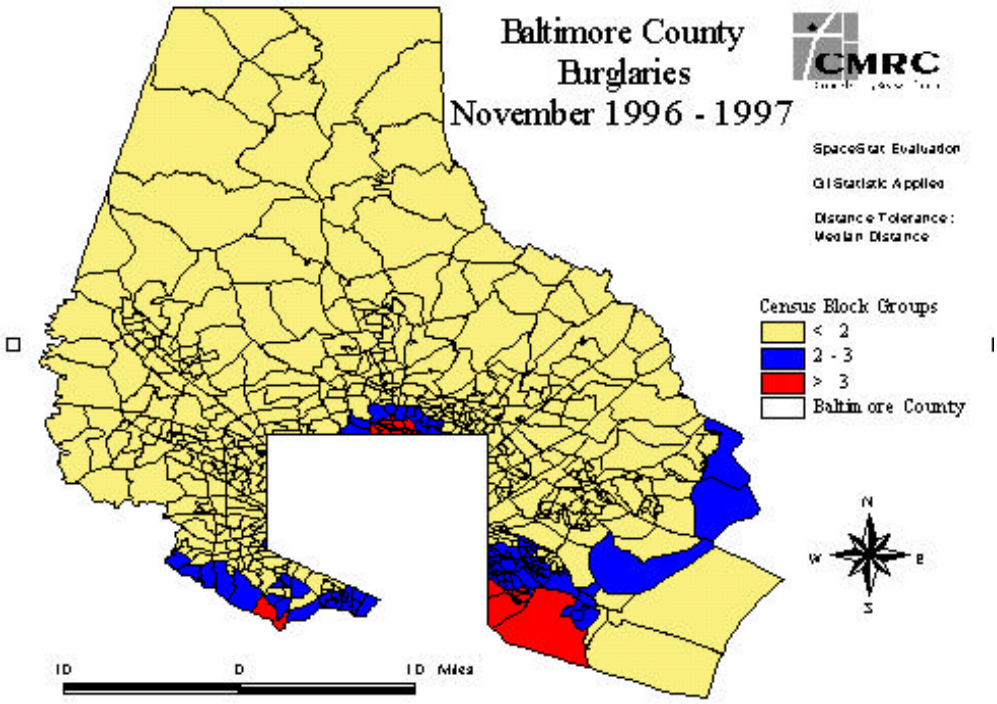
Appendix B: Robberies for Study Area 1



Appendix C: Robberies for Study Area 2



Appendix D: Burglaries for Study Area 1



Appendix E: Burglaries for Study Area 2